Journal of Digital Imaging

Part 2. Automated Change Detection and Characterization Applied to Serial MR of Brain Tumors may Detect Progression Earlier than Human Experts

Julia Patriarche, Ph.D. and Bradley Erickson, M.D., Ph.D.

An algorithm was developed which compares serial MRI brain examinations of brain tumor patients and judges them as either "stable" or "progressing". A set of 88 serial MR cases were obtained, consisting of cases which were stable and remained stable for at least 8 months, cases which were stable but progressed in less than 8 months, and cases which were progressing. The algorithm was run and its output was compared to the original clinical interpretation. Of the exam pairs which were judged stable and which remained stable at least 8 months after the later examination, the algorithm diagnosed 45/46 as stable. For exam pairs judged to be progressing, the algorithm judged 15/17 to be progressing. Of the exam pairs which were judged stable, but which went on to progress less than 8 months after the later of the pair, 16/25 were judged by the algorithm to be progressing.

KEY WORDS: Brain tumor, serial imaging, computer aided diagnosis, change detection

INTRODUCTION

The comparison of serial MR examinations of the brain is a procedure performed frequently by neuroradiologists for diseases such as tumors, multiple sclerosis, Alzheimer's disease, and others. While there are a number of objectives of such imaging, the primary objective is to detect change and characterize any changes. It was hypothesized that an automated system could be constructed, which could determine whether a brain tumor is stable or progressing, based upon the comparison of serial magnetic resonance imaging studies. A second hypothesis was proposed that this algorithm might be more sensitive to subtle changes, allowing identification of tumor progression before humans could identify progression.

METHODS

Selection of Cases

Cases were selected from a database of serial MR exams of brain tumor patients, which is held at the Radiology Informatics Lab at Mayo Clinic (the database was assembled with IRB approval). Three groups of cases were selected for this study:

1. Examination pairs which had been interpreted as showing progression. This is referred to as the "immediate progression group".
2. Examination pairs which had been interpreted as being stable, and in which follow up imaging at least 8 months after the second examination of pair continued to show no change. This is referred to as the "stable" group.
3. Examination pairs which had been interpreted as being stable, and in which follow up imaging up to 8 months or less after the second examination of the pair showed progression. This is referred to as "intermediate progression" group.

From the Department Radiology, Mayo Clinic and Foundation, Rochester, MN 55905, USA.

Correspondence to: Bradley Erickson, M.D., Ph.D., Department Radiology, Mayo Clinic and Foundation, Rochester, MN 55905, USA. e-mail: bje@mayo.edu

Online publication 29 December 2006

doi: 10.1007/s10278-006-1039-0

T1, T1 postgadolinium (Gd), and fluid attenuation by inversion recovery (FLAIR) sequences were required. The acquisition parameters were: T1: TR was between 450 and 600, TE was min full; FLAIR: TR 11000, TE 144ef; FOV was 20–22 cm. The pixel size was 0.86–0.93 mm in X and Y, and 3 mm in Z, with an interslice gap of 0 mm. All images were acquired on GE Signa 1.5 T scanners with birdcage head-coils. Images were acquired in the oblique axial plane, aligned with the anterior commissure/posterior commissure line. Cases were excluded if they contained severe artifacts, if they did not contain the entire brain in all of the above pulse sequences, or if standard registration algorithms (ITK mutual information[1] and Analyze 3D voxel based[2]) failed. Overall, five cases were excluded because of registration problems, apparently due to patient motion during the acquisition and the resultant lack of full head inclusion in all pulse sequences; and a further six cases were excluded because they did not possess each of: T1, T1 post-Gd, and FLAIR. Of the original cases, 90 comparisons remained and were used in this study. An error in file handling required the exclusion of one additional examination, leaving a total of 88 comparisons.

Change Detection Algorithm

The basic change detector algorithm used in this study has been described in a separate article (refer to companion part 1 article). The algorithm compares serial imaging studies of brain tumor patients, producing a map of change: both the nature of change (if any) and the magnitude of change for each brain voxel. In the companion article, this was

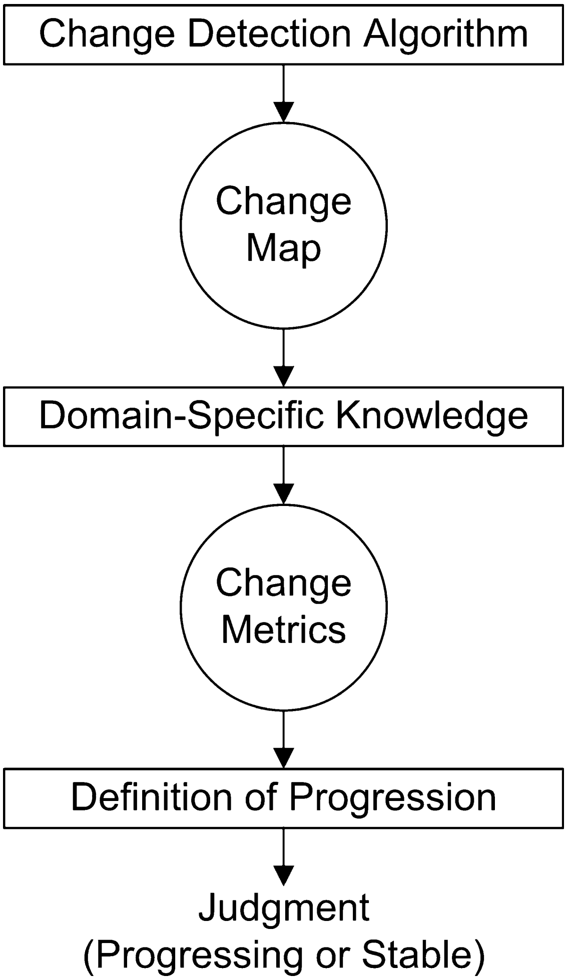
Change Detection Algorithm
Change Map
Domain-Specific Knowledge
Change Metrics
Definition of Progression
Judgment (Progressing or Stable)

Fig 1. Automated change detection process flow.

used to produce a color-coded change map superimposed on an anatomical image. In the present study, knowledge of brain tumors was applied to the change maps in the form of a series of postprocessing steps (Fig. 1), in order to generate a series of volume-wide metrics describing changes occurring within the image. The metrics described the total number of voxels/volume multiplied by the membership within each voxel, of particular types of change occurring over each image (for example, the sum of membership changes within each voxel/volume involved in a particular transition). During this process, domain-specific knowledge was introduced (middle box in Fig. 1), which attempted to incorporate information a neuroradiologist might use in analyzing images manually, and to reduce the impact of noise. Most of the specific thresholds given below reflect this knowledge, rather than a scientifically computed value. Specifically:

1. Most relevant change occurs within the vicinity of lesion, where lesion was defined in terms of minimum spatial extent and single-acquisition membership requirements, as follows. Lesion regions were required to have membership exceeding 0.25 in any pathological tissue in either the baseline or follow-up scan. Next, the resulting spatially distinct regions of pathological tissues were subjected to a requirement that at least 20 voxels (44 or 52 mm^3) in each region were required to exceed a membership threshold, where the threshold was determined on a lesion-by-lesion basis by the number of voxels in the given lesion, such that large lesions with at least 100 voxels (220 or 260 mm^3) were required to have at least 20 voxels with 0.45 membership in pathology, and could therefore be more subtle, but smaller lesions had to be more distinct, and had to possess at least 20 voxels with a threshold membership which fell along a linear scale between 0.45 up to 0.8, with smaller lesions requiring a higher membership threshold. After regions of lesion had been located, any changing voxels detected which were within 15 voxels, of lesion thus defined were considered in the computations.

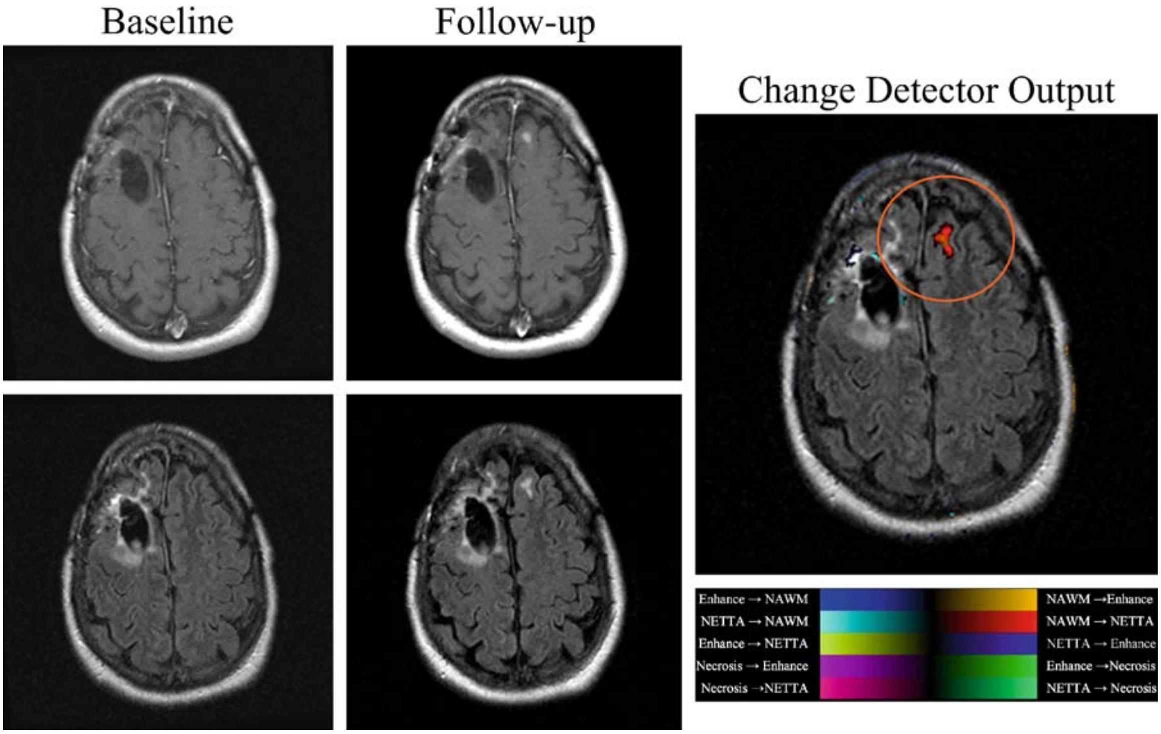

Fig 2. An example of the development of a region of satellite enhancement, the existence of which is considered sufficient grounds to state that the patient is progressing. Observe the development of enhancement from a region which was previously NAWM (colored orange), as well as the development of associated NETTA (colored red). The circle has been superimposed on the figure to highlight these features of the output map.

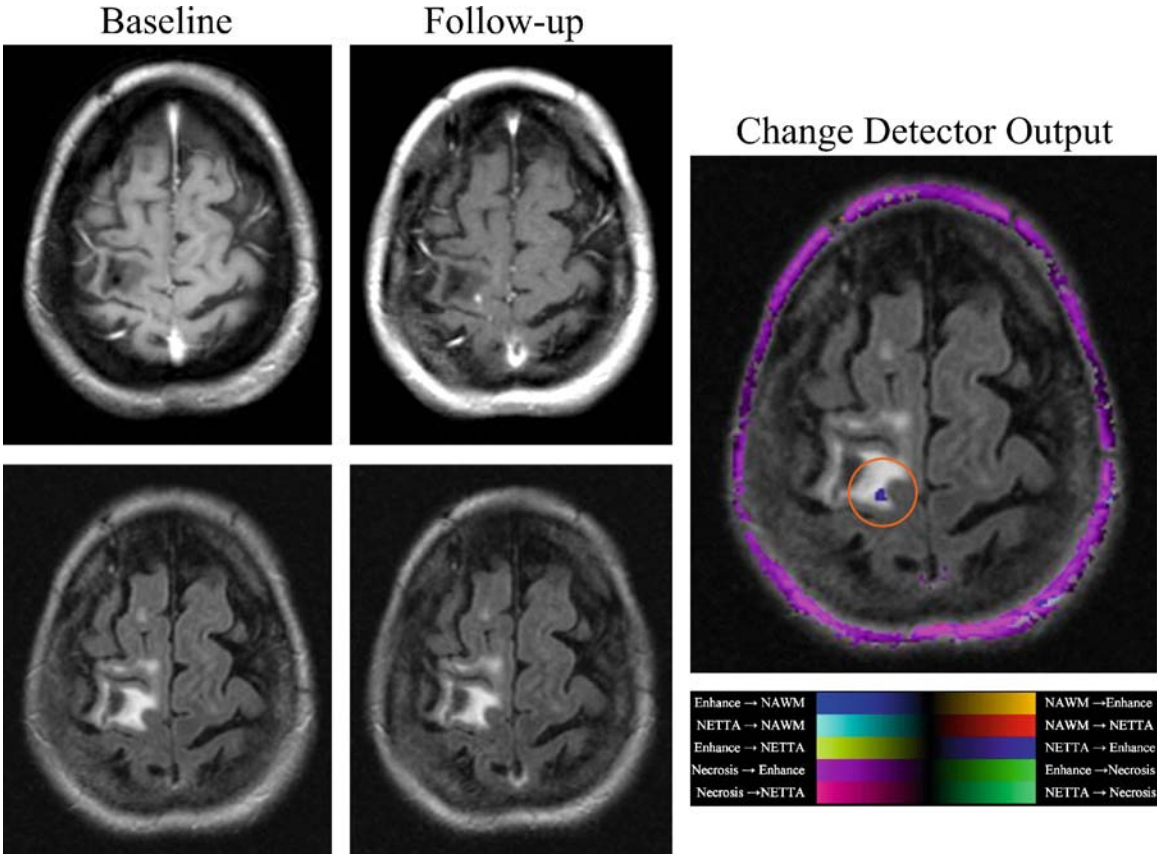

Fig 3. Another example of the development of a region of satellite enhancement. The development of enhancement from a region which was previously NETTA is colored dark blue. The circle has been superimposed on the figure to highlight this feature of the output map.

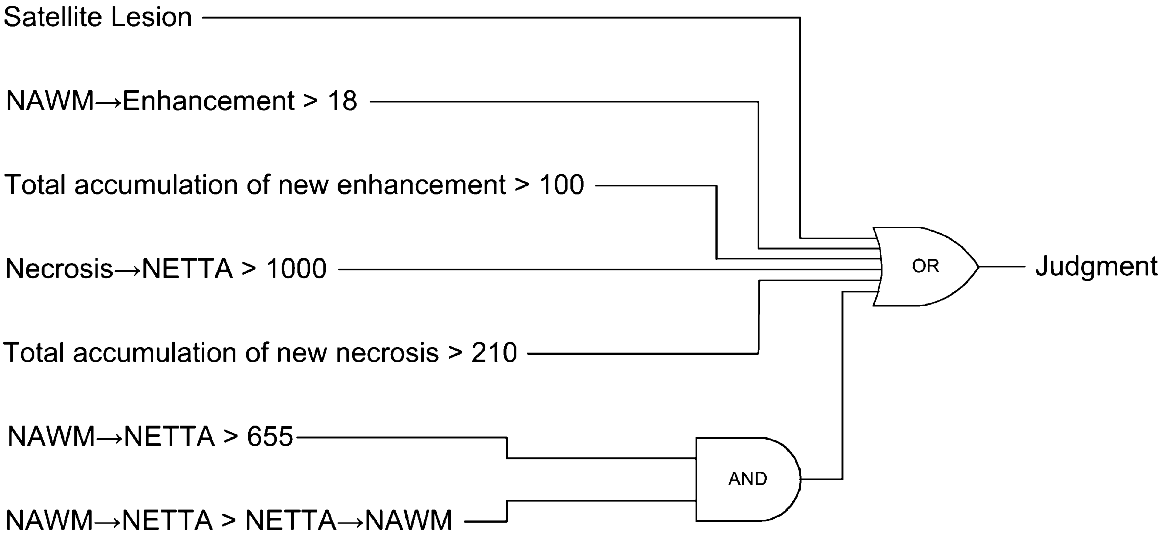

Fig 4. An expansion of the "definition of progression" step from Figure 1—the expression used to determine whether a serial comparison is stable or progressing.

2. A large group of spatially contiguous voxels (≥250 voxels, equivalent to ≥555 or 649 mm^3) containing change which were all changing in the same way were considered regardless of proximity to dominant lesion.
3. The development of a satellite region of enhancement was considered sufficient grounds to state that the patient was progressing (for examples of satellite lesions, see Figs. 2 and 3). To be considered a satellite, an accumulation of enhancement was required to be at least 15 voxels from enhancement in the baseline scan. Further, a putative satellite lesion was required to meet a minimum volume requirement (≥10 voxels, equivalent to 22 or 26 mm^3), as well as a requirement regarding the change in membership (≥0.2). In order to maximize sensitivity to small satellite lesions, while minimizing false detections, the knowledge that enhancement rarely occurs without adjacent edema was also used. A putative satellite lesion was therefore only considered if it was immediately adjacent to a region of edema greater than 100 voxels (equivalent to 222 or 259 mm^3) in extent.
4. Changes on the periphery of the brain were excluded from consideration as they are frequently due to changes in the operative site. Therefore, changes were only considered if they were at least 5 voxels interior to the brain periphery (this issue is addressed further in the discussion section, below).

The sum of membership changes was computed for each dual tissue class considered (which were NAWM↔NETTA, NAWM↔enhancement, enhancement↔NETTA, enhancement↔necrosis, and NETTA↔necrosis) in each direction (i.e., one sum was computed for NAWM→NETTA, and a separate sum for NETTA→NAWM) . Additionally, the total accumulation (i.e., regardless of which dual tissue class it originated from) of each pathological tissue (enhancement, necrosis, and NETTA) and white matter was computed. Finally, the binary value of whether or not a region of new satellite enhancement had appeared between the baseline and follow-up examinations was determined (as noted above) and considered for inclusion in the overall definition of progression.

An expression consisting primarily of a series of thresholds and logical operators (lowest box in Fig. 1) over the metrics described above was derived, to differentiate between stable and progressing cases. The equation is shown graphically in Figure 4. The symbol "→" is used to denote a voxel which is changing character from the tissue on the left of the expression, to the tissue on the right of the expression (for example, "NAWM→NETTA" means that a region of NAWM or white matter with some amount of NETTA), is acquiring an increased character of NETTA.

It should be noted that the cases used in the study were used to first validate the thresholds described above.

Analysis of Data

We divided the cases into three categories: immediate progression, intermediate progression, and stable, according to their original clinical interpretation. For each stable case, the change detector algorithm was seen as "correct" if it judged the case to be stable. For each immediate progression case, the change detector was seen as correct if it judged the case to be progressing. The two hypotheses proposed in the introduction section were assessed as follows. Hypothesis #1 was accepted as valid if the change detector algorithm correctly judged the stable and progression groups correctly at a high (>90%) rate. Hypothesis #2 was accepted as valid if hypothesis #1 was proven true, and a substantial fraction of the intermediate progression cases were found to be progressing.

RESULTS

The results for the 88 cases included in this study are shown in Table 1, which shows that the algorithm was able to distinguish stable and immediate progressing examination pairs quite

Table 1. Summary of Results

	Algorithm	
	Stable	Progressing
Stable ($N=46$)	45	1
Progressing ($N=17$)	2	15
Intermediate ($N=25$)	9	16

well—the sensitivity was 0.88, the specificity was 0.98, and the accuracy was 0.95.

The second hypothesis was also proven true: given the accuracy shown above, it appears that 16 of the 25 intermediate progression cases have changes that the change detector classified as indicating progression.

DISCUSSION

Since the advent of digital imaging, there have been thoughts of computer-automated diagnosis[3]. The work described presently represents a step towards that objective—it appears that a computer algorithm might be able to accurately compare serial MRI examinations of brain tumor patients. Of the cases which were stable over the long term, the algorithm correctly identified 45/46 as stable. The one case which was erroneously identified as progressing was identified as such due to the misinterpretation of a region of choroid plexus which became more enhancing as a new region of satellite enhancement. There are a number of ways this issue could be addressed although this is left for a future effort. Of the cases which were judged clinically to be progressing, 15/17 were identified as such by the algorithm. One of the two progressing cases which were erroneously identified as stable was identified as such because the changes were extremely small, due in part to the fact that the lesion was very small to begin with. This suggests that the absolute thresholds should instead be relative to lesion size. The other case which was erroneously identified as stable was identified as such due to an error on the part of the algorithm—a region of satellite enhancement was missed, due to its proximity to the surface of the brain. This might be addressed in ways discussed below.

Of the 25 intermediate progression cases, 16 were judged to be progressing by the algorithm. The mean time to clinically diagnosed progression (i.e., how much earlier the change detector identified progression, compared with clinical assessment) over these 16 cases was 4.1 months (σ=2 months). There are a number of possible explanations for why the algorithm apparently identified changes earlier than they were identified in the clinic. It is possible that the algorithm found changes indicating progression that could not be appreciated by a human observer; it is not unreasonable to expect to find indications of change which the algorithm might detect before such progression is evident and the neuroradiologist assigns the case a status of progression. Nine of the intermediate cases were judged to be stable by the algorithm. This also is entirely reasonable—there is no reason to believe that there are always early signs of change.

The identification of progression by the algorithm, prior to its identification by radiologists, requires further study. The algorithm may identify these changes due to greater sensitivity (the algorithm may be able to "see" things that an unassisted human cannot see or that an unassisted human misses in a particular instance), or due to operation on a different point on its receiver–operator characteristic curve compared with the neuroradiologists (the humans may be erring on the side of calling things stable).

Future Improvements

While the current algorithm has performed well within the context of the tests it was subjected to, the present study is only an initial investigation into the feasibility of automated change detection. We believe there are many avenues to improve results. One is to incorporate more knowledge. In the present study, very simple rules were used, with surprising effectiveness. When a neuroradiologist interprets a scan, however, he or she uses a great base of knowledge, and ideally so too should an automated approach. One obvious place more complex knowledge might be useful is in the identification of resection sites. In the present study, identified changes close to the surface of the brain were disregarded. A superior (though much more complex) approach would be to use knowledge of what brains look like and knowledge of what resection sites look like to identify actual resection sites. Knowledge of the way enhancement appears could also be used; specifically, the conversion of normal appearing white matter to enhancement is more suggestive of progression than is an increase in enhancement in white matter which was already enhancing. Similarly, enhancement due to postoperative changes is typically thin and linear while recurrence is more nodular.

A more sophisticated definition of progression might also improve performance. In the current study, simple thresholding, combined with logical operators, was used. Violation of any of the criteria or hard thresholds was seen as sufficient grounds for judging a case to be progressing. One obvious way to improve upon this would be to use fuzzy expressions rather than hard thresholds, i.e., development of NETTA is suggestive of progression, with more NETTA being suggestive of more progression. Development of enhancement is also suggestive of progression, with the development of more enhancement being suggestive of more progression. It would likely be best if the expression was not defuzzified into the discrete categories "stable" and "progressing" until the very last step, so that parts of the fuzzy expression could be weighed against one another as they were combined. This way, different parts of the expression could reinforce one another, i.e., development of NETTA in one region and enhancement in another, simultaneously, is more suggestive of progression than either one alone. It might even be meaningful for the output itself to be fuzzy, in lieu of the discrete categories, "stable" and "progressing". After all, all progressing tumors are certainly not the same, and there is no reason to suspect that the addition of one more voxel of NETTA should make the difference between a stable case and a progressing one (even though a definition based upon a hard threshold could yield a change of diagnosis secondary to a small increment in change such as this).

Another area for improvement could be the inclusion of good as well as bad findings in the definition of progression. The present study almost exclusively considered bad findings in its determination of status (the exception being the ability of large loss of NETTA to cause the definition of progression to disregard a smaller increase in NETTA elsewhere in the image, as shown on the bottom line of Fig. 4). Observed cues to tumor regression could be used in conjunction with the observed cues to progression, for example, development of NETTA is suggestive of progression, but loss of enhancement is suggestive of regression. How these factors should be weighed is a subject for investigation.

In the present study, the decision to base ratings on the sum of membership changes, and not upon number of changing voxels or number of spatially distinct regions of change, or other possibilities, was made arbitrarily. Additionally, the equations and thresholds used here were derived empirically. It is possible to compute the optimal factors for a given data set (perhaps with a neural network or genetic algorithm) and validate these in a larger data set.

The measures of change should probably also be normalized in some way with respect to the size of the tumor, and additionally with respect to the duration between scans. For example, it is possible that when reviewing short interval scans, smaller changes should trigger a judgment of progression compared with what would trigger such a judgment using longer interval scans. It is not obvious that this relationship should be a linear function of time, however. It remains to be established how quickly each type of pathological tissue develops. If it is established that some pathology, for example edema, develops over a very short period of time (perhaps days), then the interval between baseline and follow-up should probably not be considered. If it is determined that another aspect of pathology, for example enhancing tumor, develops over a prolonged period of time, then the duration between the baseline and follow-up scan would be very relevant. It is entirely possible that certain changes develop quickly, while others develop more slowly. In this case, the way in which the time between baseline and follow-up scans should be considered should be specific to each dual tissue class.

It should be noted that the present work is more an evaluation of principles and initial exploration, than a validation, since the same studies were used to train as to evaluate the algorithm, and there was no true gold standard. Obviously, future studies using a more rigorous approach should use separate groups for training and validation, or a leave-one-out methodology. Importantly, this study is suggestive that the quantitative metrics used appear to detect progression earlier than visual inspection. This is appealing because it supports the hypothesis that the tool could be used clinically to guide judgments; and experimentally to help in assessing, and more importantly objectively comparing new treatments. This is all in addition to providing a mechanism for the totally automated assessment of disease. It is hoped that future work with this algorithm will lead to the development of objective descriptions of disease progression and regression.

CONCLUSION

A preliminary study has been performed, showing promise for an automated change detection algorithm. Data has been generated that supports the hypothesis that this algorithm can separate cases which are stable from those which are progressing. Furthermore, it has been shown that in a substantial fraction of cases which were judged to be stable by humans but which proceeded to progression in a short period of time had changes in the images which the algorithm identified as suggesting progression. This may allow earlier detection of changes than is possible with current practice.

ACKNOWLEDGEMENT

Julia W. Patriarche was supported by NIH # NS07494-02.

REFERENCES

1. http://www.itk.org
2. Mayo Clinic: Rochester, MN, 2005
3. Lodwick G, Turner A, Lusted L, et al: Computer-aided analysis of radiographic images. J Chronic Dis 19:485–496, 1966